Fall 2020

Characterizing the Behavior of Mutated Proteins with EMCAP: the Energy Minimization Curve Analysis Pipeline

Matthew Lee
Western Washington University

Bodi Van Roy
Western Washington University

Filip Jagodzinski
Western Washington University

Characterizing the Behavior of Mutated Proteins with EMCAP: the Energy Minimization Curve Analysis Pipeline

Matthew Lee
Department of Computer Science
Western Washington University
Bellingham, United States
leem53@wwu.edu

Bodi Van Roy
Department of Computer Science
Western Washington University
Bellingham, United States
vanroyb@wwu.edu

Filip Jagodzinski
Department of Computer Science
Western Washington University
Bellingham, United States
filip.jagodzinski@wwu.edu

***Abstract*—Studies of protein mutants in wet laboratory experiments are expensive and time consuming. Computational experiments that simulate the motions of protein with amino acid substitutions can complement wet lab experiments for studying the effects of mutations. In this work we present a computational pipeline that performs exhaustive single-point amino acid substitutions *in silico*. We perform energy minimization as part of molecular dynamics (MD) of our generated mutant proteins, and the wild type, and log the energy potentials for each step of the simulations. We motivate several metrics that rely on the energy minimization curves of the wild type and mutant, to explore quantitatively the effects of the mutations. Two case studies are discussed and analyzed to showcase the utility of our approach to identify the least and most impactful mutations.**

***Index Terms*—Energy minimization, mutation, computational biochemistry, structural biology**

I. Introduction

The structure of a protein is determined by the sequence of amino acids that comprise it, and how they interact with each other and the solvent. Changing a single amino acid can change the stability and global structure of the protein and its substructures drastically. Performing exhaustive *in vivo* or *in situ* laboratory experiments on physical proteins to determine the outcome of any given mutation is infeasible, as the process of mutating a bacteria strain, growing the culture, lysing the cells, purifying the protein and then studying the resulting protein, is prohibitively resource- and time-intensive. Synthetically producing the protein *in vitro* is similarly difficult or otherwise time-consuming.

The need for a practical computational alternative has given rise to various methods for generating mutations *in silico*. To refine protein structures that are generated *in silico*, energy minimization as part of Molecular Dynamics minimizes the potential energy of the protein by performing many slight alterations of the biomolecule, which seeks to a find an energetically feasible conformation.

In this work, we use our custom software to perform all possible amino acid substitutions at a specific residue in a protein, followed by energy minimization of the wild type and generated protein mutant structures. We generate exhaustively all substitutions at a single location so that we may identify those protein variants that behave very differently from all other mutants. Identifying the most impactful substitutions is useful information that has relevance to drug design studies and inferring the effects of debilitating diseases. We motivate and present two metrics, *energy turbulence* and deviation from the energy trendline, that we refer to a *disparity*, to provide insights into the effects of the mutations. Our approach is not dependent on large data sets of structural or thermodynamic properties of mutants, nor does it require vast compute resources that are common for approaches that rely on machine learning (ML) based methods.

II. Related Work

Wet lab experiments provide the gold standard for directly measuring the effects of mutations on a protein's stability. Matthews et al. have studied many mutants of Lysozyme from Bacteriophage T4 [1], [3], [14]. Unfortunately such wet lab studies are time consuming and often expensive, requiring specialized hardware and materials.

Computational approaches have been developed over the years to predict the effects of mutations on protein structure and stability. Early work relied on heuristic energy functions, molecular dynamics simulations, and rotamer libraries [7], [10], [17]. In our previous work we relied on rigidity analysis to assess via a rigidity analysis approach how mutating to glycine destabilizes a protein's structure [8]. Machine learning, which is a branch of artificial intelligence, that involves algorithms that learn to make predictions from data, has also extensively been used to infer the effects of mutations. These efforts range from the use of Support Vector Machines [4], to the use of Random Forest approaches [13], to our own various efforts which relied on support vector machines, random forests, and deep learning methods [5].

All of these, and related works, require access to data sets of experimental or computed structural of biochemical features of proteins and their mutants, as for example the ProTherm and ProNIT databases that detail thermodynamic properties of proteins and protein-nucleic acid interactions [9]. The machine

learning-based approaches also often require vast amounts of compute resources, in part because they are statistics-based.

Understanding the effects of amino acid substitutions also has implications in docking and related work. For example docking approaches have been used for predicting how amino acid substitutions affect protein complexes, as is the case with REMC [21]. Similarly, there are efforts underway to assess how mutations affect protein-drug complexes [20]. Some other approaches still use free model techniques to describe protein behavior (EdaRose [19]), while template based approaches rely on homology modeling [12]. Certain computational methods are more effective at predicting the effects of certain mutations, as has been the case with computational alanine scanning in the industry for years [15].

In contrast, EMCAP is meant to be generalized to study the substitution of any residue in a protein, without relying on web lab data nor needing vast compute resources for machine learning approaches.

III. Materials and Methods

To infer which amino acid substitutions at a specific residue location result in mutants that behave very differently from all other mutants, EMCAP utilizes multiple steps. Mutants are generated *in silico*, an off-the-shelf molecular dynamics engine performs energy minimization of the wild type and mutants, and several custom metrics are calculated from the energy minimization log files to help identify which substitutions are the most impactful. The entire compute pipeline is self contained and is able to run on a single core computer, in near real time.

EMCAP exhaustively generates all single amino acid substitutions at a single residue in a protein using ProMute [2]. It is able to generate the structure files for all 19 mutants for a single residue location in minutes, using a single-core computer. The NAMD molecular dynamics [16] engine is used to perform the short runs of molecular dynamics. For this work, a total of 500 steps of energy minimization are performed, using an explicit solvent model.

A. Generating Protein Mutants

The first step of EMCAP, the MultiMutant step (Fig. 1), generates the protein mutants. It invokes ProMute, followed by a brief energy minimization run for each mutant. At each time step of the energy minimization, the total energy of the protein — wild type or mutant — as calculated by NAMD, is retained. The energy minimization step serves to account for any steric clashes that might have been introduced during the mutation step, and also serves to create an energy log for making an energy minimization curve that is later analyzed to identify the most impactful substitutions. MultiMutant collects the energy minimization logs of all the mutants.

B. Aggregation of Energy Curves for Mutants

The numerical analysis portion of EMCAP includes multiple sub-steps (Fig. 2). First, EMCAP collects all energy minimization data for the specified mutations into a single file. For each energy step, EMCAP compares the energy of each mutant to the energy of the wild type. The energy disparity at every time step between a mutant and the wild type is recorded. It is these disparity values, as well as the raw energetics properties at each time step of the wild type and mutant, that are analyzed.

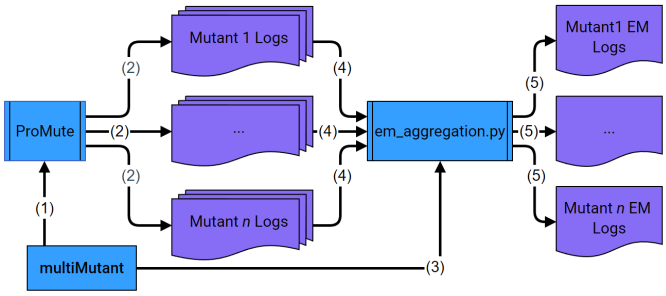


Fig. 1: MultiMutant, the first component of EMCAP. MultiMutant invokes ProMute (1), which generates mutants *in silico* (2), whose last step is energy minimization (EM). MultiMutant invokes em_aggregation (3), which analyzes the mutant log files (4) to generate the EM processed logs (5).

C. Quantifying Mutant Protein Qualities

EMCAP implements several metrics to help distinguish which amino acid substitutions, based on their energy minimization profiles of the mutants and the wild type, are most unlike the other amino acid substitutions. Disparity is the main measurement from which the other metrics are derived. Disparity refers to the cumulative difference between the energy potentials of a mutant protein and its wild type, measured at each time step of energy minimization. The two disparity measurements we found most useful are 1) Disparity as Percentage of WT Total Potential, and 2) Weighted Disparity as Percentage of WT Total Potential. Representing these disparities as percentages helps contextualize their significance and thus serves as a better comparative measure when looking

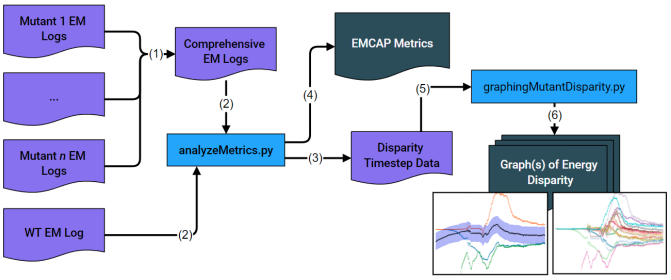


Fig. 2: The analysis and graphing capabilities of EMCAP. The individual mutants' EM log files are collated (1), and compared to the EM log file of the wild type (2). AnalyzeMetrics tallies energy disparity metrics for each time step in the EM log files (3), and outputs the turbulence and divergence metrics (4). graphingMutantDisparity processed the timestep data (5) and generates graphs of energy trendlines and outliers (8).

at different proteins. There are also two ancillary themes, *divergence* and *turbulence*. Turbulence is our metric to measure the chaotic behavior of a mutant during the course of an energy minimization, and divergence describes how much a given mutant diverges from the general trendline of all other mutants generated by mutating the wild type at that residue. The shaded regions in (c) and (d) of Figs. 4 and 6 are ± 0.5 standard deviations. They indicate outliers and are not for the purposes of calculating divergence.

1) Absolute Disparity, AD: This is the cumulative sum of disparities between a mutated protein's energy potential and its wild type's energy at each time step of energy minimization:

$$AD = \sum_{n=s}^{500} |EP_M(n) - EP_{WT}(n)| \tag{1}$$

where n is the current time step of energy minimization, s is the number of initial timesteps that are ignored to account for the random seed of atom velocities and vectors in the initial stages of energy minimization ($s = 6$ for all the results presented here), while $EP_M(n)$ and $EP_{WT}(n)$ are the energy potential of the mutant and wild type at time step n.

2) Disparity Ratio, DR: This is the AD and total energy potential of the wild type, as a % of the wild type's energy:

$$DR = 100 * \left(\frac{AD + |TP_{WT}|}{|TP_{WT}|} - 1 \right) \tag{2}$$

where TP_{WT} is the total, cumulative energy potential of the wild type over the course of the energy minimization.

3) Weighted Disparity, WD: This determines the significance of energy disparity between the mutant and wild type based on the time step where it occurs, with higher importance given to earlier time steps. A weighted figure is calculated at every step of energy minimization, and summed:

$$WD = \sum_{n=s}^{500} \frac{1000 - n}{500} * |EP_M(n) - EP_{WT}(n)| \tag{3}$$

4) Weighted Disparity Ratio, WDR: This is the WD and TP_{WT}, taken as a percentage of TP_{WT}:

$$WDR = 100 * \left(\frac{WD + |TP_{WT}|}{|TP_{WT}|} - 1 \right) \tag{4}$$

IV. Results and Discussion

To demonstrate the usefulness of EMCAP, we studied the mutations in the crystal structure of PDZ1 complexed with APC (PDB 3RL7) [22], and mutations in PDZ2 (PDB 2BYG) [6]. PDZ1 and PDZ2 are classes of PDZ domains, which bind the extreme C-terminus of target proteins and are critical components of signaling and trafficking pathways [11]. We choose PDZ domains specifically because of their importance in protein interactions, and thus knowing which mutations are most impactful can provide insights to guide mutational studies to better understand these proteins.

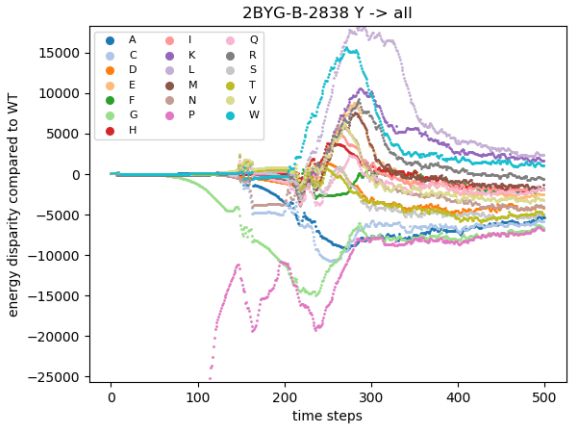


Fig. 3: 2BYG mutations on chain B, residue 2838.

A. PDZ Structure 2BYG

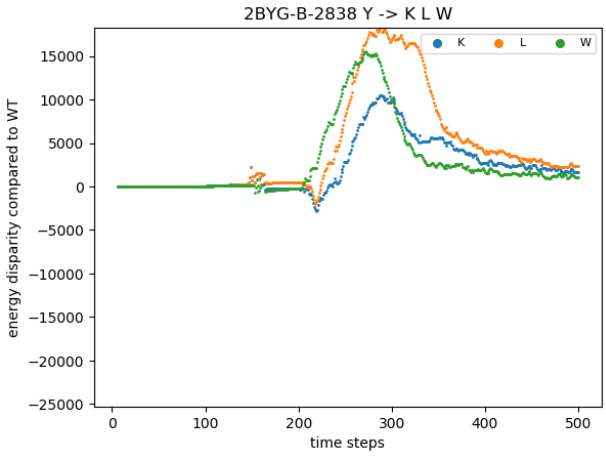
If one were to look at the energy minimization curves for all mutants at residue 2838 for 2BYG (Figure 3), it would not be possible to easily discern with substitutions to that residue have the biggest effect on the PDZ domain to be able to recover from the mutation. However, inspecting the tabular results for the mutations to residue 2838 (Table I) reveals that mutation to residues KLWGPL have the most impact on the PDZ domain, while residues ACGIVQ have the least impact. The disparity and turbulence plots for those residues are shown in Figure 4.

B. PDZ structure 3RL7

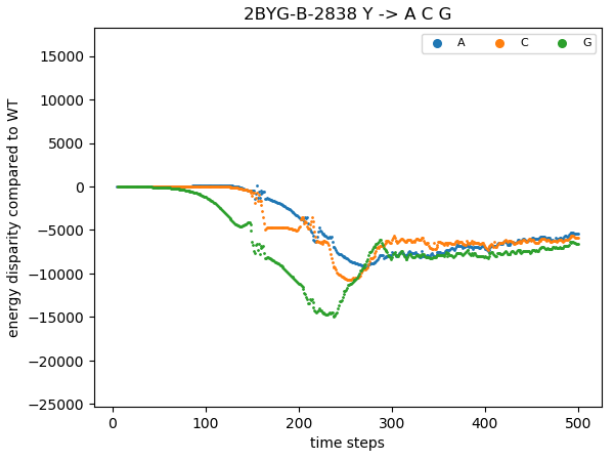
Similar to 2BYG, inspecting the energy minimization curves for 3RL7 (Figure 5) does not reveal which substitutions have the biggest effects on the PDZ domain to be able to recover

TM	DR	WDR	TO	DO	TG	SD	MED
I	1.31	1.72	201	33	D2	1 898.86	1 033.34
H	1.05	1.36	284	36	A1	1 567.81	831.52
K	3.18	4.25	338	277	A1	3 010.30	2 507.17
M	1.49	2.05	165	55	D2	2 240.08	1 178.66
L	5.75	7.77	323	303	D2	5 781.69	4 537.19
N	2.76	3.66	49	39	B	4 270.56	2 177.13
A	5.58	7.38	4	207	D1	9 398.77	4 405.03
C	5.76	7.71	34	208	C1	9 651.77	4 546.62
E	1.97	2.61	171	70	A2	2 884.56	1 556.17
D	2.52	3.14	98	42	A2	4 340.01	1 989.49
G	8.34	11.66	28	361	C2	13 756.85	6 583.37
F	0.94	1.18	130	42	D3	1 720.25	739.55
Q	1.45	1.90	52	32	B	2 398.59	1 145.53
P	28.06	48.47	69	486	C1	50 406.77	22 152.33
S	2.96	3.67	121	63	B	5 121.86	2 336.14
R	1.74	2.47	230	82	A1	2 362.11	1 370.27
T	2.85	3.52	111	56	B	4 914.00	2 253.22
W	3.69	5.18	294	301	D3	4 307.53	2 913.77
V	2.01	2.59	205	35	D1	2 942.76	1 584.31

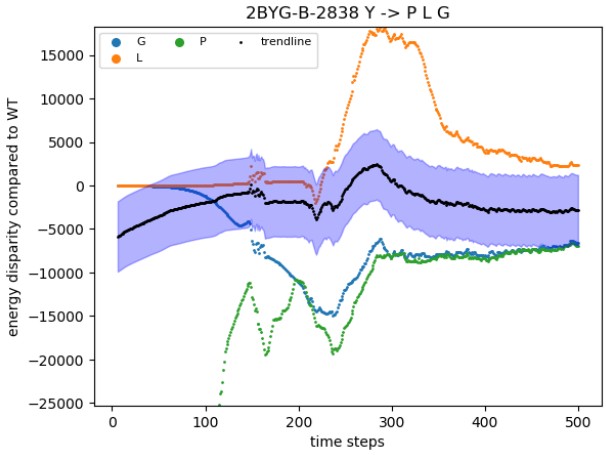
TABLE I: Mutations to 2BYG, chain B, residue 2838; wild type=Y. TM=Target Mutation, TO=Turbulent Outliers, DO=Divergent Outliers, TG=Target Group, MED=Mean Energy Disparity, DR=Disparity Ratio, WDR=Weighted Disparity Ratio, SD=Standard Deviation of energy values.



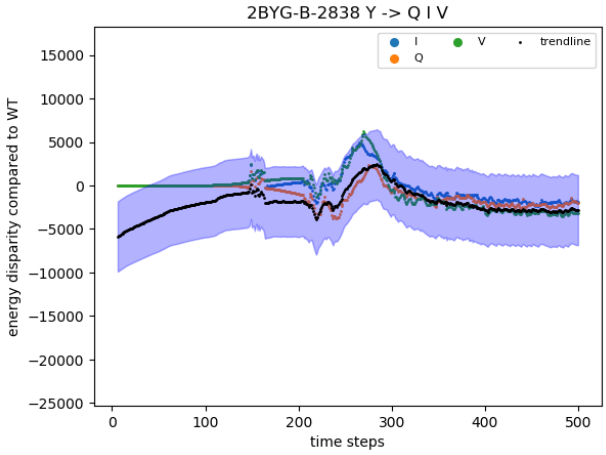
(a) most outliers, energy disparity



(b) fewest outliers, energy disparity



(c) most outliers from general trendline



(d) fewest outliers from general trendline

Fig. 4: Mutations to 2BYG residue 2838 chain B. The 3 mutations with the most (a) and fewest (b) outliers on their energy disparity curves. The 3 mutations with the most (c), and fewest (d) outliers from the general trendline in their disparity curves. The shaded purple regions represent ± 0.5 standard deviation from the trendline.

from the mutation. However, inspecting the tabular results for the mutations to residue 2837 (Table II) reveals that mutation to residues FMQDL have the most impact on the PDZ domain, while residues EPTKRN have the least impact. The disparity and turbulence plots for those residues are shown in Figure 6.

C. Comparative Utility of Different Metrics

To explore which metrics offer a unique perspective about the effect of an amino acid substitution, we show the top 5 disparity ratios by magnitude, top 5 weighted disparity ratios by magnitude, top 5 outliers by casual visual inspection of the graphed disparity curves, turbulence as described by outliers from a mutation curve's mean, and divergence from the trendline as described by outliers from that trendline, for 2BYG and 3RL7 in Table III and Table IV. Proline, which is known to be very disruptive when introduced via an amino acid substitution [18], is identified by several of the metrics as a very impactful substitution. However, at the 5^{th} row of

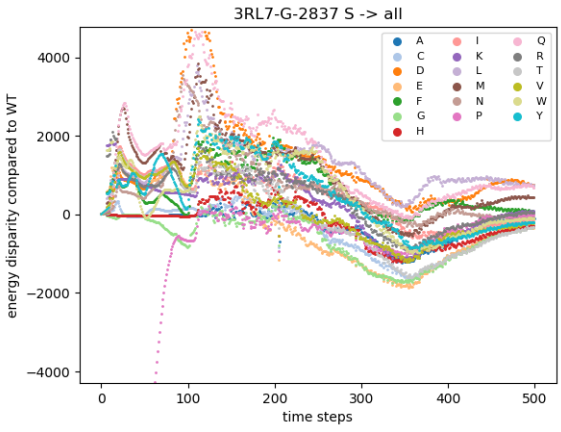
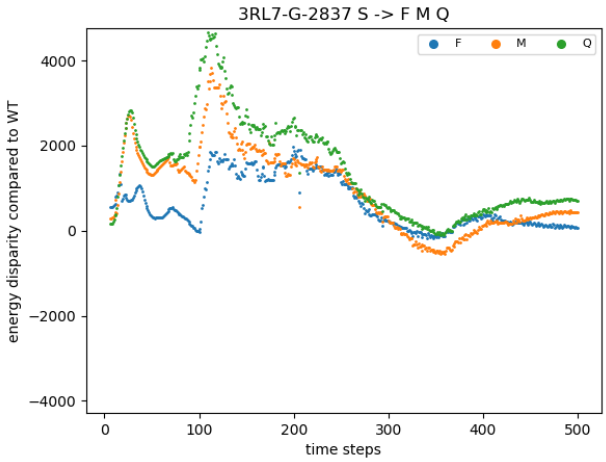
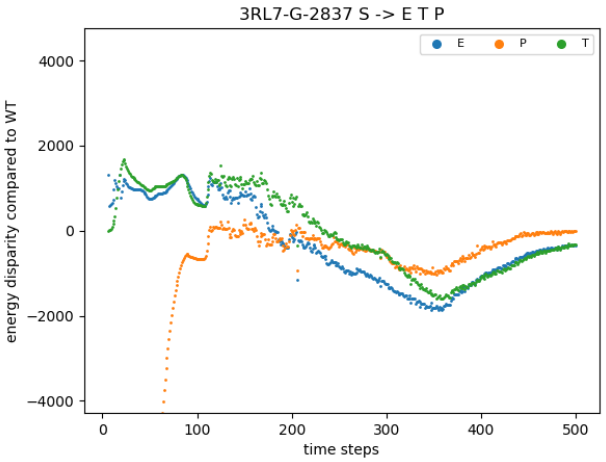


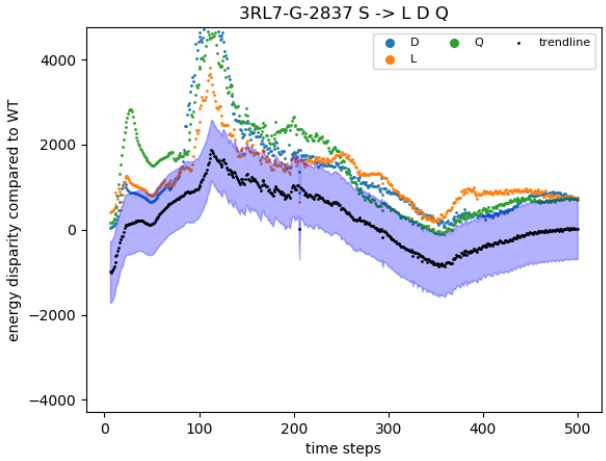
Fig. 5: 3RL7 mutations on chain G, residue 2837.



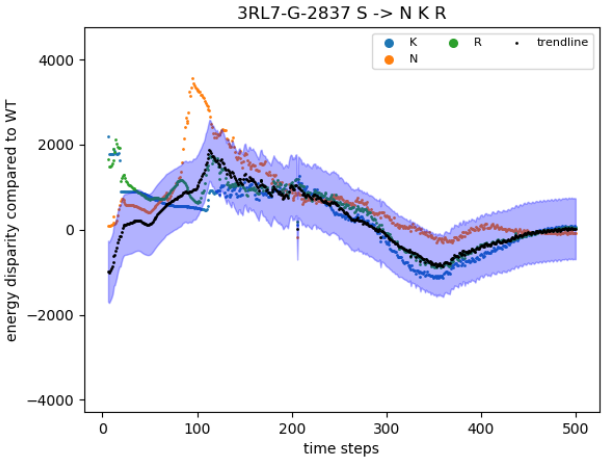
(a) most outliers, energy disparity



(b) fewest outliers, energy disparity



(c) most outliers from general trendline



(d) fewest outliers from general trendline

Fig. 6: Mutations to 3RL7 residue 2837 chain G. The 3 mutations with the most (a) and fewest (b) outliers on their energy disparity curves. The 3 mutations with the most (c), and fewest (d) outliers from the general trendline in their disparity curves. The shaded purple regions represent ± 0.5 standard deviation from the trendline.

each of those tables, there is very little consensus, indicating that either the mutations to those two PDZ domains have very different effects, or that the different metrics are able to discern between biophysical effects that the mutations can have on the PDZ domains and their interaction targets. We leave to discern which of these is true, to future work.

V. Future Work

We envision several next steps for this work. Firstly, the current version of EMCAP does not permit a user to fine-tune the parameters for calculating the disparity and turbulence metrics. Allowing such fine-tuning might permit a user to discern better between the least- and most-impactful mutations if upon a first run, all mutations produce minimization curves that are visually indistinguishable. Although that didn't happen in our case studies of 3RL7 and 2BYG, it is possible. For example, changing the standard deviation value might allow to determine outliers easier that diverge from the energy minimization trendline for all mutations. Or, changing the factor (0.5, 1.0, 1.5, etc.) of standard deviation mentioned in subsection C of section III might change the general type of disparity curves that are ranked as highly divergent. For example, consider two mutations, one with a disparity curve that exhibits brief, wild divergences from the wild type in a few places along the minimization curve, but otherwise closely adheres to the trendline, and another with a disparity curve that mirrors the trendline almost exactly but does so with a constant offset on the y-axis. Discriminating by a low factor of standard deviation would rank the second mutation as more divergent than the first, as most of the points on the second line would be considered outliers while only a few points on the first line would fall into that category. However, discriminating by a sufficiently high factor of standard deviation would reverse the ranking, as the second line would fall entirely with the

TM	DR	WDR	TO	DO	TG	SD	MED
G	7.97	10.76	91	348	C2	1 353.92	623.58
F	8.48	13.79	411	127	D3	608.54	663.58
E	11.55	17.17	49	273	A2	1 536.12	903.74
D	17.07	27.54	255	409	A2	1 173.23	1 335.41
C	6.88	9.21	90	265	C1	1 111.03	538.46
A	4.03	5.33	91	152	D1	702.79	315.11
N	9.23	15.59	306	50	B	866.00	721.52
M	13.74	22.85	365	201	D2	970.30	1 074.79
L	15.93	24.89	339	386	D2	644.91	1 245.82
K	8.31	13.17	128	43	A1	832.96	650.11
I	8.56	13.92	262	61	D2	686.39	669.12
H	5.69	7.71	93	91	A1	848.74	445.26
W	11.57	18.46	275	88	D3	1 017.49	904.92
V	8.78	13.87	129	59	D1	970.19	686.81
T	10.84	16.43	64	119	B	1 274.92	847.99
R	8.79	14.34	187	30	A1	766.99	687.30
Q	18.36	30.18	402	420	B	1 130.50	1 436.03
P	25.70	47.70	54	258	C1	6 289.72	2 010.38
Y	11.10	17.74	244	80	D3	982.34	868.04

TABLE II: Mutations on 3RL7, chain G, residue 2837; wild type=S. TM=Target Mutation, TO=Turbulent Outliers, DO=Divergent Outliers, TG=Target Group, MED=Mean Energy Disparity, DR=Disparity Ratio, WDR=Weighted Disparity Ratio, SD=Standard Deviation of energy values.

DR	WDR	Visual Inspection	Turbulence	Divergence
P	P	P	K	P
G	G	L	L	G
C	L	G	W	L
L	C	W	H	W
A	A	C	R	K

TABLE III: EMCAP analysis of 2BYG, chain B, residue 2838. The ranking of the 5 highest (highest at top) values for each of DR=Disparity Ratio, WDR=Weighted Disparity Ratio, Turbulence, and Divergence.

standard deviation and would not be considered an outlier, but the spikes in disparity on the first line would still contain outliers.

DR	WDR	Visual Inspection	Turbulence	Divergence
P	P	P	F	Q
Q	Q	D	Q	D
D	D	Q	M	L
L	L	L	L	G
M	M	G	N	E

TABLE IV: EMCAP analysis of 3RL7, chain G, residue 2837. The ranking of the 5 highest (highest at top) values for each of DR=Disparity Ratio, WDR=Weighted Disparity Ratio, Turbulence, and Divergence.

REFERENCES

[1] Tom Alber, Sun Dao-Pin, Keith Wilson, Joan A Wozniak, Sean P Cook, and Brian W Matthews. Contributions of hydrogen bonds of thr 157 to the thermodynamic stability of phage t4 lysozyme. *Nature*, 330(6143):41–46, 1987.

[2] Erik Andersson and Filip Jagodzinski. Promuteht: A high throughput compute pipeline for generating protein mutants in silico. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 655–660, 2017.

[3] Jeffrey A Bell, Wayne J Becktel, Uwe Sauer, Walter A Baase, and Brian W Matthews. Dissection of helix capping in t4 lysozyme by structural and thermodynamic analysis of six amino acid substitutions at thr 59. *Biochemistry*, 31(14):3590–3596, 1992.

[4] Jianlin Cheng, Arlo Randall, and Pierre Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 62(4):1125–1132, 2006.

[5] Ramin Dehghanpoor, Evan Ricks, Katie Hursh, Sarah Gunderson, Roshanak Farhoodi, Nurit Haspel, Brian Hutchinson, and Filip Jagodzinski. Predicting the effect of single and multiple mutations on protein structural stability. *Molecules*, 23(2):251, 2018.

[6] Jonathan M Elkins, Evangelos Papagrigoriou, Georgina Berridge, Xiaowen Yang, Claire Phillips, Carina Gileadi, Pavel Savitsky, and Declan A Doyle. Structure of pick1 and other pdz domains obtained with the help of self-binding c-terminal extensions. *Protein science*, 16(4):683–694, 2007.

[7] Dimitri Gilis and Marianne Rooman. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *Journal of molecular biology*, 272(2):276–290, 1997.

[8] Filip Jagodzinski, Jeanne Hardy, and Ileana Streinu. Using rigidity analysis to probe mutation-induced structural changes in proteins. *Journal of bioinformatics and computational biology*, 10(03):1242010, 2012.

[9] MD Shaji Kumar, K Abdulla Bava, M Michael Gromiha, Ponraj Prabakaran, Koji Kitajima, Hatsuho Uedaira, and Akinori Sarai. Protherm and pronit: thermodynamic databases for proteins and protein–nucleic acid interactions. *Nucleic acids research*, 34(suppl_1):D204–D206, 2006.

[10] Christopher Lee and Michael Levitt. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature*, 352(6334):448–451, 1991.

[11] Ho-Jin Lee and Jie J Zheng. Pdz domains and their binding partners: structure, specificity, and modification. *Cell communication and Signaling*, 8(1):1–18, 2010.

[12] Marc F Lensink, Sameer Velankar, Andriy Kryshtafovych, Shen-You Huang, Dina Schneidman-Duhovny, Andrej Sali, Joan Segura, Narcis Fernandez-Fuentes, Shruthi Viswanath, Ron Elber, et al. Prediction of homoprotein and heteroprotein complexes by protein docking and template-based modeling: A casp-capri experiment. *Proteins: Structure, Function, and Bioinformatics*, 84:323–348, 2016.

[13] Yunqi Li and Jianwen Fang. Prots-rf: a robust model for predicting mutation-induced protein stability changes. 2012.

[14] Blaine HM Mooers, Walter A Baase, Jonathan W Wray, and Brian W Matthews. Contributions of all 20 amino acids at site 96 to the stability and structure of t4 lysozyme. *Protein Science*, 18(5):871–880, 2009.

[15] Irina S Moreira, Pedro A Fernandes, and Maria J Ramos. Computational alanine scanning mutagenesis—an improved methodological approach. *Journal of Computational Chemistry*, 28(3):644–654, 2007.

[16] James C Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of computational chemistry*, 26(16):1781–1802, 2005.

[17] Martine Prevost, Shoshana J Wodak, Bruce Tidor, and Martin Karplus. Contribution of the hydrophobic effect to protein stability: analysis based on simulations of the ile-96—-ala mutation in barnase. *Proceedings of the National Academy of Sciences*, 88(23):10880–10884, 1991.

[18] Ken N Savage and John M Gosline. The effect of proline on the network structure of major ampullate silks as inferred from their mechanical and optical properties. *Journal of Experimental Biology*, 211(12):1937–1947, 2008.

[19] David Simoncini, Thomas Schiex, and Kam YJ Zhang. Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 85(5):852–858, 2017.

[20] Edward Thompson, Tess Thackray, Cecilia Kalthoff, Ryan Rapoport, and Filip Jagodzinski. Using energy-minimization profiles to measure protein resistance to drugs. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–6, 2020.

[21] Zhe Zhang, Uwe Ehmann, and Martin Zacharias. Monte carlo replica-exchange based ensemble docking of protein conformations. *Proteins: Structure, Function, and Bioinformatics*, 85(5):924–937, 2017.

[22] Zhenyi Zhang, Hua Li, Leyi Chen, Xingyu Lu, Jian Zhang, Ping Xu, Kui Lin, and Geng Wu. Molecular basis for the recognition of adenomatous polyposis coli by the discs large 1 protein. *PloS one*, 6(8):e23507, 2011.